Retrieval-Augmented Style Transfer

Chen-En Ma, Jiayi Shen, Hongwei Liao
Brown University
Providence, RI, United States
chen-en_ma@brown.edu, jiayi_shen1@brown.edu, Hongwei_liao@brown.edu

I. Introduction

Style transfer (Fig 1) has long been a cornerstone of creative AI, enabling the transformation of an image's visual appearance based on a desired artistic or stylistic reference. However, the success of style transfer often hinges on the availability of appropriate style images, which may not always be readily accessible. To address this limitation, we propose Retrieval-Augmented Style Transfer (Fig 3), a novel framework that combines retrieval-based augmentation (Fig 2) with diffusion models to achieve efficient and flexible style transfer.

Our approach begins by leveraging natural language prompts to retrieve relevant style images from a vast database, ensuring that users can efficiently access stylistic references tailored to their needs. These retrieved style images are then processed by enhanced diffusion models, which seamlessly integrate the retrieved styles into the target image. By incorporating retrieval-augmented generation, our method eliminates the need for manually curated style images, broadening the scope of creative applications and reducing the barrier to high-quality style transfer.

This project demonstrates how retrieval-augmented workflows can extend the capabilities of generative AI, offering a practical and scalable solution for generating styled content across a variety of artistic domains.

Fig. 1. Style Transfer; transferring styles from the style image to the base image

II. Related Work

Our work integrates advancements from several domains, including multimodal representation learning, retrieval-augmented generation, style transfer, and diffusion models and enhancement. Below, we discuss prior works that inform and contextualize our approach.

A. Multimodal Representation Learning

Multimodal retrieval plays a crucial role in our approach by aligning representations of text and image modalities. Contrastive Predictive Coding [1] utilized unsupervised learning techniques to learn meaningful representations by predicting future data points in a latent space. CLIP[2] pioneered multimodal alignment by training on large-scale image-text pairs, enabling strong cross-modal retrieval capabilities. Our method builds on this foundation by combining multimodal representation with a retrieval mechanism to ensure that the retrieved images are both semantically relevant and stylistically aligned with the input prompt.

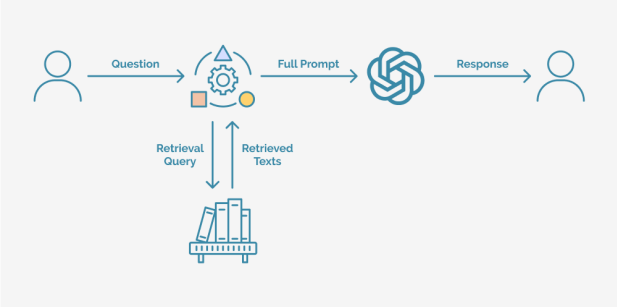

Fig. 2. Retrieval-Augmented Generation; combining retrieved content with the input prompt for generation models

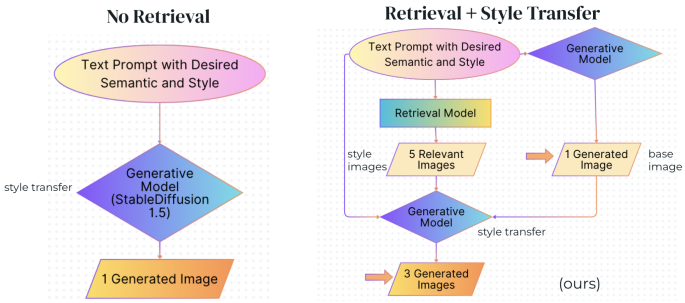

Fig. 3. No Retrieval v.s. Retrieval-Augmented Style Transfer; Instead of one-step generation based on prompt, we first retrieve relevant style images to guide and improve model generation.

B. Retrieval-augmented Generation

Retrieval-augmented generation has been explored primarily in text-based knowledge-intensive tasks[3, 4]. For image generation, Instance-conditioned GANs[5] leveraged additional attributes to guide image synthesis. Re-Imagen[6] uses a retrieval-based approach to improve image generation, especially for rare or unseen entities. Our work extends these ideas

to the domain of text-to-image synthesis by using retrieval to guide the style transfer process.

C. Style Transfer

The concept of style transfer was popularized by a neural algorithm introduced in 2016, which leverages feature representations from convolutional neural networks to transfer artistic style[7]. More recently, StyleGAN-NADA[8] demonstrated the effectiveness of a CLIP-guided method for adapting generative models to new style domains. StyleDiffusion[9] explicitly extracts the content information and implicitly learns the complementary style information, yielding interpretable and controllable content-style disentanglement and style transfer. These works lay the groundwork for integrating style transfer into retrieval-augmented frameworks.

D. Diffusion Models and Enhancement

Diffusion-based generative models have shown remarkable progress in text-to-image synthesis. Stable Diffusion and Imagen[10, 11] demonstrated state-of-the-art results in generating high-quality, semantically accurate images. Retrieval-augmented diffusion models[12] further improve text-to-image generation by incorporating external information into the synthesis process. IP-Adapter[13] introduced a lightweight module that enhances text-to-image diffusion models by integrating image prompts without retraining the base model. ControlNet[14] enhanced the controllability of text-to-image diffusion models by incorporating additional conditioning information. DreamBooth[15] allows users to personalize text-to-image diffusion models by training them on a few images of a specific subject, enabling highly customized image generation. Inspired by this line of research, our work utilizes retrieval to enhance the style fidelity and diversity of images generated via diffusion models.

Our method is the first work that combines retrieval-augmented generation and diffusion-based style transfer into a unified framework. By using multimodal retrieval to identify semantically and stylistically relevant images, we enhance a base generative model with retrieval-guided style augmentation. This approach surpasses traditional text-to-image pipelines that rely solely on prompt engineering or pre-trained models.

III. Dataset

We selected the SemArt [16] for our retrieval-augmented style transfer task because of its rich, multi-modal nature, combining fine-art paintings with artistic descriptions. This multi-modal structure is ideal for training models to understand both visual and textual information, crucial for style transfer. The dataset's artistic diversity offers a wide range of stylistic features such as color, texture, and patterns, which help the model transfer and blend multiple styles effectively.

Additionally, SemArt (Fig 4) is designed for semantic art understanding, making it particularly valuable for learning how different styles interact on both a visual and conceptual level. Its multi-modal retrieval capabilities further enhance our approach, allowing for the retrieval of relevant style images based on text prompts. This is essential for guiding the diffusion model during the generation phase, ensuring the final image reflects the desired stylistic influences.

Fig. 4. SemArt is an artistic dataset designed for multi-modal retrieval.

IV. Methodology

This project introduces a retrieval-augmented technique to improve the style transfer capabilities of generative models, specifically focusing on ControlNet-enhanced diffusion architectures. As Fig 3 shows, our proposed method leverages Stable Diffusion to generate a base image guided by a prompt describing the desired content and style. The same prompt is then used to retrieve a set of five relevant style images from a curated dataset. These retrieved images, combined with the generated base image and original prompt, are fed into the enhanced diffusion model to produce three final outputs that better align with the desired stylistic and semantic attributes.

This approach integrates retrieval and generative methods to bridge the gap between purely text-based and visual reference-based style transfer, leading to more consistent and high-quality results. The system's ability to incorporate external style references ensures that the outputs are rich in diversity while maintaining fidelity to the input prompts and retrieved stylistic cues.

A. Text-to-Image Retrieval Training and Inference Pipeline

This approach leverages transfer learning to train a representation model for text-to-image retrieval with significantly reduced computational costs, as shown in Fig 5. The methodology employs a frozen CLIP model, comprising pre-trained vision and text encoders, alongside a lightweight projection model shared between modalities. By freezing the CLIP backbone, the approach minimizes the number of trainable parameters by 99% compared to training the entire model from scratch.

The training process incorporates contrastive learning via the InfoNCE Loss [1] function, which minimizes the distance between matching text-image pairs in the representation space while maximizing the distance between mismatched pairs. This ensures robust alignment between text and image embeddings, fostering effective retrieval performance.

Once trained, the model extracts text and image representations and utilizes a retriever powered by a sentence transformer's [17] search and re-ranking capabilities. This retriever identifies the top 5 relevant style images for any given prompt, providing a critical input for downstream style transfer (Fig 6). The approach achieves competitive performance (Fig 9) while maintaining computational efficiency, making it well-suited for scalable real-world applications.

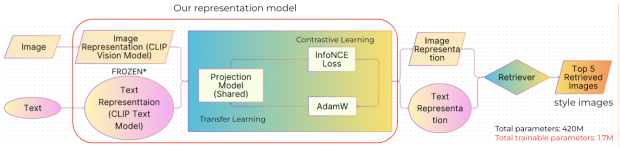

Fig. 5. Text-to-Image Retrieval Training and Inference; By leveraging transfer learning, we trained the model only on 1.7M parameters, instead of 420M.

Fig. 6. Snippet of Retrieval; based on a prompt, our retrieval system will encode the prompt and search through an image representation array to find the most relevant style images. The ground truth image is at the second place among top 5.

B. Retrieval-Augmented Style Transfer with Diffusion Models

In our retrieval-augmented style transfer, we compare two methods (Fig 7) for enhancing Stable Diffusion: one using ControlNet and the other using DreamBooth. The first method leverages ControlNet, a technique that allows for fine-grained control over image generation by guiding the diffusion model with additional structured information. The second method utilizes DreamBooth, which fine-tunes Stable Diffusion to learn specific concepts or styles, enabling it to generate images with more personalized or stylized features. Both methods integrate with our retrieval-augmented framework to guide style transfer, offering different strengths in terms of control and customization.

1) Controlnet + Stable Diffusion: ControlNet enhances Stable Diffusion by introducing a mechanism for conditioning the image generation process on structured auxiliary inputs, such as edge maps, poses, or other visual features (Fig 8). In the context of style transfer using multiple style images, ControlNet can play a critical role in guiding the diffusion model to harmonize and apply style attributes from multiple references. Details of how it works include:

1) **Encoding Style Features:** The multiple style images are processed to extract key stylistic features, such as color schemes, textures, and patterns. These features can be converted into structured representations, such as feature maps or latent embeddings.
2) **Conditioning with ControlNet:** ControlNet accepts these structured inputs alongside the base text prompt or image to act as a guide for the generation process. By embedding the extracted style features into the conditioning mechanism, the model is directed to incorporate specific stylistic elements into the output image.
3) **Dynamic Fusion of Styles:** ControlNet's architecture allows it to interpret and blend the features from multiple style images dynamically. This enables the model to integrate elements from each style image cohesively, rather than simply mimicking a single reference.
4) **Fine-Grained Control:** With ControlNet, users can exert granular control over how each style is applied. For example, a user could specify that one style's textures dominate the background while another style's color palette defines the subject.

By augmenting Stable Diffusion with ControlNet, the model gains the ability to synthesize visually compelling images that reflect a blend of multiple style references, enabling complex and nuanced style transfer tasks.

2) DreamBooth + Stable Diffusion: In this work, we propose a novel approach to experiment with the capability of DreamBooth for style transfer, even though DreamBooth was not originally designed for this purpose. According to its paper, DreamBooth is primarily intended for embedding personalized objects or subjects into images and synthesizing them in diverse scenes while preserving their key features. However, we adapt its methodology to explore style transfer by replacing the concept of a personalized object with that of a stylistic concept.

To achieve this, we modify the instance prompt from its original format, such as "A photo of a <T> dog," to "A photo of {our_text_prompt} in <sks> style." In this case, the model learns the <sks> token as representing the style derived from input images. We also incorporate the class-specific prior preservation loss to leverage the semantic prior embedded in the model. The class prompt used for prior preservation is "a photo of an artwork." During inference, the same prompt format as the instance prompt is used, ensuring consistency between training and fine-tuning.

This approach allows DreamBooth, when fine-tuned with Stable Diffusion, to learn and represent stylistic concepts in a manner analogous to how it embeds personalized subjects. By treating style as the target concept, the model is able to generate outputs that reflect the stylistic features of the training images. Our experimental results demonstrate that this adaptation of DreamBooth opens new possibilities for exploring its use in style transfer tasks.

V. Results

We divide this section into two part: 1) retrieval and 2) style transfer with ControlNet and Dreambooth. In retrieval evaluation, we use Accuracy@5 and Mean Reciprocal Rank

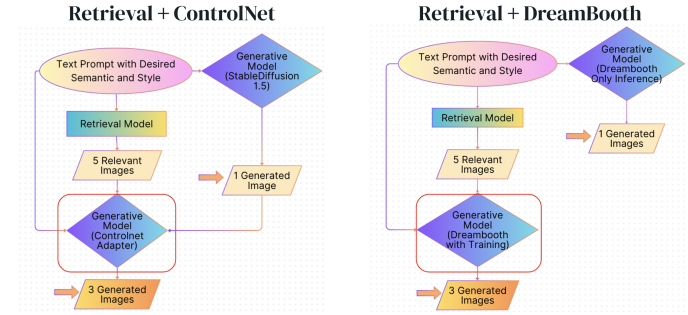

Fig. 7. ControlNet v.s. DreamBooth; The only difference lies in the last generation model for the transfer output. One uses ControlNet-enhanced Stable Diffusion, while the other uses DreamBooth-enhanced Stable Diffusion.

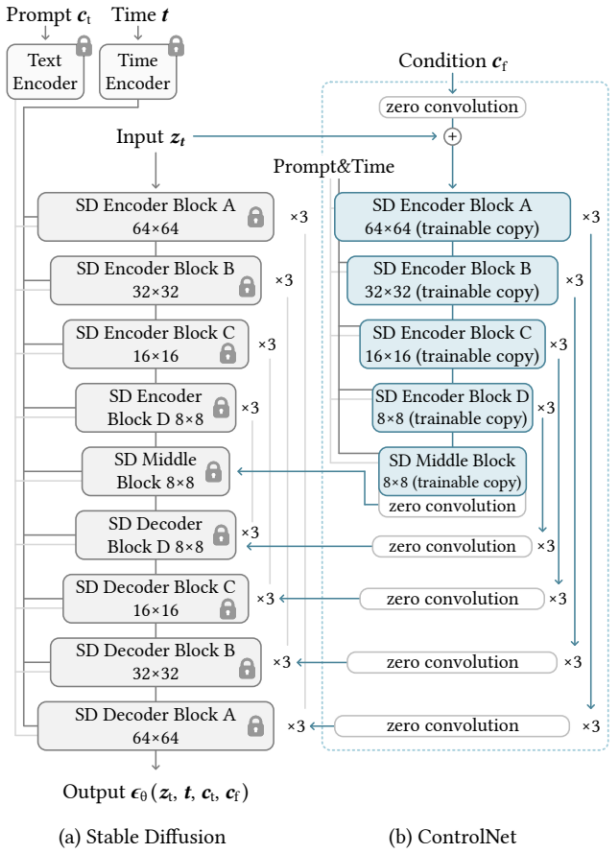

Fig. 8. Stable Diffusion's U-net architecture connected with a ControlNet on the encoder blocks and middle block. The locked, gray blocks show the structure of Stable Diffusion V1.5 (or V2.1, as they use the same U-net architecture). The trainable blue blocks and the white zero convolution layers are added to build a ControlNet.

(mRR) as the evaluation metrics. For style transfer with ControlNet and Dreambooth, we borrowed the idea from CLIP-based scores[15, 18] to quantitatively evaluate the average similarity between style images and generated image. We also presents qualitative analysis with a few style transfer results.

A. Retrieval

As shown in Fig. 9, by employing transfer learning and contrastive learning on the CLIP model, our retrieval system achieved an Accuracy@5 of **0.91** and an mRR of **0.81**. These scores represent significant improvements when compared with the baseline performance of zero-shot CLIP. Specifically, the Accuracy@5 score increased by **59%**, indicating a notable enhancement in identifying relevant images within the top five results, while the mRR score surged by an impressive **268%**, reflecting a much stronger ability to rank the ground truth image closer to the top.

These improvements highlight the impact of fine-tuning through transfer and contrastive learning in aligning multi-modal representations more effectively. The results suggest that the proposed enhancements not only improve accuracy but also prioritize relevance more efficiently, making the system more robust for real-world retrieval tasks.

	Accuracy @ 5 ↑	MRR ↑
CLIP Zero-Shot	0.57	0.22
CLIP w/ Transfer Learning	**0.91**	**0.81**

Fig. 9. Retrieval Results. Accuracy@5 measures the percentage of cases where the ground truth image appears in the top 5 retrieved results for a given text input. MRR measures how high the ground truth image appears in the ranked results. Compared with the results using zero-shot CLIP, the Accuracy@5 score increases 59%, and mRR score increases 268%.

B. Style Transfer: ControlNet VS DreamBooth

1) Division of Results: We divided the generated images into four sets.

- **Set One:** contains only 1 image generated by ControlNet approach given only the text prompt, which is the same as the text prompt we used for retrieval;
- **Set Two:** contains only 1 image generated by DreamBooth approach given only text prompt;
- **Set Three:** contains 3 images generated by ControlNet approach given text and five retrieved images as inputs, and the retrieved images are used for style transfer purpose;
- **Set Four:** contains 3 images generated by DreamBooth approach given text and five retrieved images.

2) Qualitative Results - Images: As shown in Figs. 10, 11, and 12, at the top are the five retrieved images for style reference, below is the image grid made for visual comparison.

- **The first row** are generated images with only text input,
- **The second row** are images with text and five retrieved images as inputs, and we choose the best result among 3 generated images for visualization purpose;
- **The first column** shows results from using ControlNet approach,
- **The second column** shows results from Dreambooth approach.

Overall, the second column has a more matched style with retrived images, compared to the first column, so our innovation on retrieve-augmentation shows its effect.

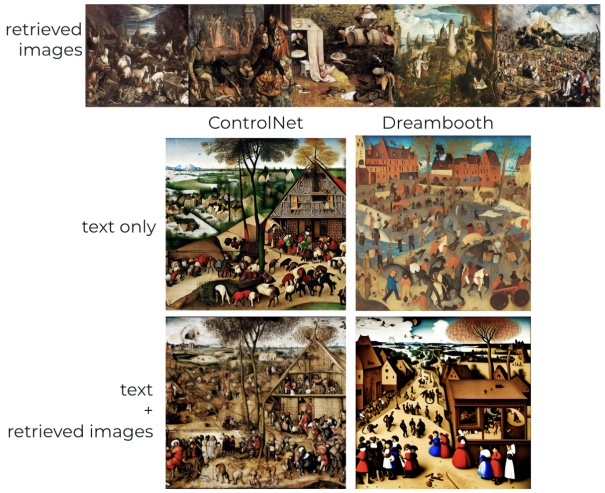

Fig. 10. Image Results 1. Text prompt: Create an artwork of a joyful trip with Northern Renaissance-style inspired by the compositions of Pieter Bruegel the Elder, reinterpreted by Baltens. Emphasize intricate detail, narrative complexity, and vibrant scenes of everyday life. Incorporate dynamic groupings and a rich, textured atmosphere, reflecting the influence of Bruegel's 1564 painting while showcasing Baltens' distinct artistic vision and interpretation.

Fig. 11. Image Results 2. Text prompt: Create an artwork of company meeting with people wearing black suits, using Renaissance-style depiction inspired by Leonardo da Vinci's works. Emphasize central perspective and symmetry. Incorporate dynamic compositions, vibrant colors, and symbolic details that reflect Leonardo's innovative interpretation of the Gospel narrative.

3) Quantitative Results - CLIP Img2Img Similarity Score: Inspired by the evaluation metric CLIP-I in IP-Adapter [13] and DreamBooth [15], our CLIP Img2Img Similarity Score calculates the average pairwise cosine similarity between the CLIP [2] embeddings of generated images and their corresponding retrieved images. The results are summarized in Fig. 13, comparing the ControlNet and DreamBooth approaches under two input configurations: using text prompts only and using text prompts combined with five retrieved images for style transfer.

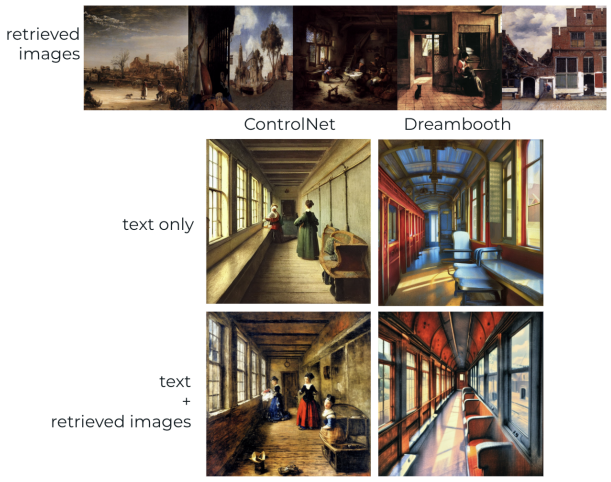

Fig. 12. Image Result 3. Text prompt: Create an artwork of train carriage with Dutch Golden Age-inspired scene in the style of Pieter de Hooch, blending still-life qualities with precise perspective, and luminous colors. Capture a serene, frozen moment imbued with timelessness, where the interplay of light and structure evokes a sense of quiet eternity.

The ControlNet approach consistently outperforms DreamBooth in both configurations. Specifically, when using only the text prompt, the ControlNet achieves a similarity score of 0.669, compared to DreamBooth's 0.542. This gap widens when style information is introduced through five retrieved images, where the ControlNet reaches a similarity score of **0.728**, while DreamBooth achieves only 0.533. These results indicate that the ControlNet approach demonstrates better style alignment between generated and retrieved images, particularly when leveraging additional style references, underscoring its robustness in capturing stylistic nuances.

	ControlNet	Dreambooth
Text Only	0.669	0.542
Text + 5 Retrieved Image	**0.728**	0.533

Fig. 13. CLIP Img2Img Similarity Score

4) Limitations: Despite the promising results, there are several limitations in the style transfer step due to the inherent differences in each generative approach:

- Firstly, although the ControlNet approach achieves higher performance than DreamBooth, it heavily relies on its "base image" — the generated image produced from only the text prompt. Since the final output of the ControlNet approach applies learned style modifications to this base image, any semantic mismatch between the base image and the text prompt will lead to unreasonable outputs, regardless of the quality of the style transfer. Additionally, the human-driven process of selecting a suitable base image can become tedious and time-consuming when scaled to larger datasets. On the other hand, even though

DreamBooth doesn't require picking the "base image," it's designed to be used for incorporating personalized item in to generated image. Using DreamBooth approach for style transfer is mostly an experiment.

- Secondly, while CLIP-I has been widely adopted in prior works [15] [13] [18], it is not guaranteed to be the most appropriate evaluation metric for image-to-image style similarity. Defining "style" quantitatively remains a challenging task, and there is currently no universal consensus on the optimal metric for this purpose.
- Lastly, although DreamBooth does not outperform ControlNet in terms of the CLIP Img2Img Similarity Score, it demonstrates greater diversity in its generated images. For instance, in Fig. 11, when given the text input "people wearing black suits" without specifying gender, DreamBooth generates an image of male-looking individuals in modern-style black suits. After being trained on the retrieved images, it produces an image of three female-looking individuals wearing black-and-white dresses, likely influenced by the abundance of dress-wearing figures in the retrieved images. Similarly, in Fig. 12, the ControlNet outputs resemble the bottom-right image in Fig. 11, consistently depicting indoor scenes of 2-3 individuals in dresses. This suggests potential entanglement between specific words in the text prompt and image features, caused by the underlying stable diffusion model. Conversely, DreamBooth's results in Fig. 12 exhibit a more innovative alignment with the text prompt, capturing the essence of a "train carriage" and evoking a "frozen moment imbued with timelessness, where the interplay of light and structure evokes a sense of quiet eternity."

VI. Ethics

When integrating style transfer, there are important societal and cultural implications to consider. A key concern is the creative ownership of artists, as style transfer involves incorporating art elements from various artists and cultures, potentially leading to ownership issues and unacknowledged contributions. Additionally, diffusion models may unintentionally generate harmful or inappropriate outputs, reinforcing stereotypes or spreading misinformation. Style transfer also blurs the line between personal creativity and ownership, as it allows users to quickly learn and generate unique works, which may impact artists' creativity. Major stakeholders include artists, machine learning engineers, and users. Artists may lose control over their intellectual property, while users might unknowingly infringe copyrights, and engineers could face legal consequences if protected images are used in training datasets.

VII. Division of Labor

Chen-En Ma (cma72):

1) Co-proposed and co-designed the project based on past experience and observation of current research
2) Identified suitable and available datasets for our tasks
3) Trained and evaluated the representation model in the fashion of transfer learning and contrastive learning
4) Identified and adapted off-the-shelf retrieval functions for our tasks
5) Completed the full retrieval-augmented style transfer pipeline with ControlNet on which team members can test with different generative models

Jiayi Shen (jshen95):

1) Co-proposed and co-designed the project based on past experience and observation of current research
2) Implemented the pipeline with dreambooth, trained the model and generated images based on our retrieved images.
3) Set up the CLIP image-to-image similarity score pipeline to evaluate the similarity of the style between the generated images and the retrieved images.

Hongwei Liao (hliao13):

1) Co-proposed and co-designed the project based on past experience and observation of current research
2) read through some different related research papers
3) tried out some text-to-image and image-to-image generative models

VIII. Challenges

We were planing to train our diffusion-based style transfer models. However, after a few trials and research investigating, we found it challenging due to their high computational demands, requiring significant hardware resources and large, diverse datasets. The iterative denoising process can lead to training instability and slow convergence, while the complex model architectures and pipelines demand significant expertise to implement and optimize. Balancing style fidelity with content preservation is difficult, as is achieving generalization across diverse styles without overfitting. Additionally, memory limitations and the need for subjective evaluation of results further complicate the training process, making it resource-intensive and time-consuming.

IX. Reflection

We are generally satisfied with how our project turned out, especially when it can given us some surprising results compared to normal style transfer models. Our base goal was to demonstrate a pipeline that retrieves relevant style images form a given text prompt and the uses these references to guide style transfer. We successfully met this goal, by implementing a retrieval model that works efficiently with a large dataset and integrating the retrieved images into a pre-trained stable diffusion model.

This project basically works out the way we expected it to. Despite that we could not train our own style transfer model, this project demonstrates the success of combining retrieval with off-the-self diffusion-based style transfer methods. Our retrieval method can efficiently collect relevant style images based on prompt and thus provides precise information for diffusion models to complete the task. Our projects also presents

the effectiveness of using ControlNet and DreamBooth as enhancement methods to guide diffusion models in image generation.

Initially, we planned on training our style transfer model. Over time, we recognized that we do not have enough time and computing resources for this process. Another shift occurred when we discovered that ControlNet can helps stable diffusion model to handle multiple images as input. We pivoted from using a fine tuned stable diffusion model to implementing our own functions to enabling normal stable diffusion to be able to handle multiple images and text prompts.

At the end of this project, we were satisfied with the pipeline we implemented, but if we had more time, we would like to train our own style transfer model for better image generation, and continue fine-tune our retrieval model for better representations. We would also like to use a larger dataset to reduce data scarcity issue in our current database of the retrieval model.

Finally, we would like to mention several key takeaways from this project:

1) **Efficient Representation Learning:** We observed with great interest that transfer learning has the ability to significantly reduce the overall number of trainable parameters. By leveraging this reduction, we minimized the need for extensive computational resources during model training. Additionally, transfer learning substantially decreased the time required for training, enabling the model to produce highly expressive representations that capture the relations between text and images within a matter of minutes.
2) **Diffusion Model Enhancement:** We are surprised that, with the enhancements provided by ControlNet and DreamBooth, diffusion models can be optimized to adapt themselves to domain-specific prompts and generation tasks. ControlNet allows the incorporation of additional conditioning inputs, improving the model's control over generation outputs. DreamBooth facilitates personalized fine-tuning by leveraging limited domain-specific data. These methods not only enhance the adaptability of diffusion models but also expand their applicability to artistic fields.
3) **Limitations for the Project:** Despite promising results, we also learn the limitations of our generative approaches in style transfer. While ControlNet delivers high-quality, it also relies heavily on the base image. Ill base images could sabotage the overall results even if good style prompt and style images are provided. On the other hand, though DreamBooth offers great diversity in the images it generates, it limits the optimal prompt to a certain format, which could be less intuitive and natural for user to follow. Moreover, DreamBooth suffers from the overhead of fine-tuning, making it less efficient when adapting to new tasks.

As the project reaches its conclusion, we are thrilled and proud to share the achievements we've accomplished over the past few months. This journey has been a testament to our dedication, collaboration, and hard work, and we are excited to celebrate the milestones we've reached. The progress we have made is not just a reflection of our team's efforts, but also a stepping stone toward even greater successes in the future of deep learning.

REFERENCES

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2020.

[4] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," 2022. [Online]. Available: https://arxiv.org/abs/2112.04426

[5] A. Casanova, M. Careil, J. Verbeek, M. Drozdzal, and A. Romero-Soriano, "Instance-conditioned gan," 2021. [Online]. Available: https://arxiv.org/abs/2109.05070

[6] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," 2022. [Online]. Available: https://arxiv.org/abs/2209.14491

[7] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423.

[8] R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," 2021. [Online]. Available: https://arxiv.org/abs/2108.00946

[9] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2211.13203

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[11] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion

models with deep language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[12] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://arxiv.org/abs/2204.11824

[13] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2308.06721

[14] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[15] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2022.

[16] N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," 2018.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[18] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2104.08718